Empirical Modeling in Economics

Specification and Evaluation

Clive W.J. Granger

Professor of Economics

University of California, San Diego

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge, CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1999

Printed in the United Kingdom at the University Press, Cambridge

Typeset in 10.5/16pt Meridien in QuarkXPress® [SE]

A catalogue record for this book is available from the British Library

ISBN 0 521 66208 7 hardback
ISBN 0 521 77825 5 paperback

Contents

CHAPTER ONE

The specification of empirical models

Models and changes

On May 25, 1961 President John F. Kennedy issued a challenge to the scientists and engineers of the United States to put an American on the moon by the end of the decade. The first moon landing took place on July 20, 1969. To meet the challenge many research problems had to be solved in rocketry, control theory, materials science, and in other fields. This was successfully completed in the appropriate sequence, within the dead-line and, I believe, within the cost constraints imposed by Congress. This is an example of a successful challenge. It had substantial funding attached and the researchers embraced it with some enthusiasm, clearly believing that the objective was achievable. In contrast, on May 18, 1997, President William J. Clinton challenged the US health community to find a vaccine to prevent AIDS within the next decade but he did not promise extra funding. The response was muted, some saying that the challenge was impossible, some that it would be better to try to find a cure for AIDS rather than a preventative vaccine and there has been little discussion of this challenge since.

How would economists respond to a similar challenge? Economists in a country could be asked to attempt to solve, or at least reduce, some particular economic problem. Possible examples of such problems would be if there were small regions that had clearly lower income levels or growth rates or if some groups in the economy faced particularly high unemployment rates. The challenge to the economists would be to alleviate these perceived difficulties within a given horizon and for an agreed cost. A different challenge might be to consider how to make efficient use of senior citizens, aged say 60 to 75, who are mentally active, are less able physically, but who do not wish to fully retire. In another example, economists could be asked to find economic incentives supplied by the richer countries to persuade Brazil to reduce deforestation in its rain forests. Given some funding, could the economists in the nation(s) organize themselves to tackle the problem, specify what the problem is in a precise form and state what would comprise a solution that would be satisfactory to the challenger? I have been talking to colleagues about this possibility and they are not optimistic. A common position is that if the economists ever did find a solution it would not be acceptable to other major policy-making groups in society with their own viewpoints, such as politicians, lawyers, sociologists, and journalists, and so the economist's solution would not be accepted by society unless it was a Pareto optimum. However this viewpoint can be ameliorated by using second-best solutions and large doses of bargaining and game theory which economists should be good at.

Some possible challenges are difficult to define. Suppose that the economists are asked to substantially reduce poverty but that the government statisticians define poverty as those falling into the lowest 12 percent of the income distribution. In this case poverty can never be solved as there will always be 12 percent at the bottom of whatever income distribution occurs. An alternative definition of poverty may be arbitrary in some other way.

There have been two minor challenges to economists in recent years of which I am aware. In the EC econometricians have been encouraged to study whether there is evidence for convergence of regional economies towards some aggregate level. The question is obviously of interest to politicians in a community of countries about to adopt a common currency. The difficulty with this research has been in deciding on the correct definition of convergence. A number of definitions have been used, some of which are quite unsatisfactory. In the US economists have been provided with extra funds to study the economics of global warming. For several years Congress has provided these funds to the National Science Foundation which were then allocated using competitive bids. Unfortunately there was no clear objective for the research and so it has been dissipated over many topics rather than being focused in just a few directions. However, the main example that I will be using in this lecture is based on a global warming project. Overall, I think that economists and their techniques when evaluated and compared to results from other fields have performed in a bimodal way; sometimes rather well, sometimes badly. This may be because the challenges are so poorly defined.

To tackle a practical problem an economist will need to build an empirical model. That is, a model of the actual economy or at least that part of it that is relevant to the problem. The main topic covered in this chapter is the process of building such a model. That is, its specification, interpretation, cost, what it is not, and how it varies according to its objectives. In my second chapter I will discuss the important questions concerning the evaluation of models: How do we know if a model is any good and how that will depend on the objective of the modeling exercise.

A relevant starting point is to admit that there is no single clearly best way of approaching the question of how to specify an empirical model. Virtually every econometrician and applied economist has their own way and we can each point out weaknesses in the approach used by others. The result is that several different models are produced rather than just one and I view this as a strength in the situation rather than a weakness as it means that we have different models to compare and then we learn by making comparisons. As new data arrive it is easier to hill-climb to better models using several starting points than just a single one.

Eventually it will be necessary to concentrate on just a few of the models and drop the rest, but this is best done after a proper evaluation exercise, preferably using a post or out of sample data set.

There are several reasons for the lack of agreement between empirical modelers but a major one is the huge complexity of a modern economy. In the US, for example, the population of over 260 million contains about 100 million family units all making many economic decisions hourly, daily, weekly, monthly, annually, or longer depending on the type of decision. Only a very small percentage of the economic outcomes from these decisions are recorded, such as purchases, hours worked, and investments made. Of these only an extremely small percentage are made public. Government and some private agencies collect these data, aggregate them, possibly seasonally adjust them, and then make them generally available. Even with standard data sets there is too much information to use in the typical model. It is thus necessary to begin by selecting variables to be used and finding the corresponding available data, perhaps after using both temporal and cross-sectional aggregation. All this implies that the eventual model will certainly be an approximation to the actual economy, but probably a very crude approximation. Because of this inevitable reduction process, from the many decisions that make up the actual economy to the empirical model, and the many possible ways to form an approximation, it is obvious why various alternative models can occur. Clearly at some point in the modeling process some selection process will have to be applied, as part of the evaluation procedure.

Before proceeding I need to make clear a piece of terminology. In what follows I will restrict the use of the word "model" just to empirical models based on data from an actual economy. Such models will often be distinguished from information in the form of what may be called theory. In this simplistic viewpoint a theory starts with a set of consistent assumptions and then produces logical consequences from them in a form relevant to economic questions. On some occasions this theory is best expressed using very sophisticated mathematics, "best" here meaning the most rigorous and compact although not necessarily the easiest to understand. To have something easier to use and to interpret a simple version of the theory can be formed, an approximation, and this is sometimes called a theoretical model. However, I will call all such constructs just "theory." I will take the attitude that a piece of theory has only intellectual interest until it has been validated in some way against alternative theories and using actual data.

A sculptor once said that the way he viewed his art was that he took a large block of stone and just chipped it away until he revealed the sculpture that was hidden inside it. Some empirical modelers view their task in a similar fashion starting with a mass of data and slowly discarding them to get at a correct representation. My perception is quite different. I think of a modeler as starting with some disparate pieces – some wood, a few bricks, some nails, and so forth – and attempting to build an object for which he (or she) has only a very inadequate plan, or theory. The modeler can look at related constructs and can use institutional information and will eventually arrive at an approximation of the object that they are trying to represent, perhaps after several attempts. Model building will be a team effort with inputs from theorists, econometricians, local statisticians familiar with the data, and economists aware of local facts or relevant institutional constraints. As projects get large, the importance of team-work becomes emphasized.

The Amazon project

I will illustrate many of the problems faced when undertaking an empirical study by using parts of a study of the dynamics of deforestation in the Amazon region of Brazil. The analysis attempted to consider how the rate of change of deforestation is influenced by economic, demographic, and policy changes.

It is useful to start with some basic facts. The Amazon rain forest is the largest remaining uncleared forest in the world covering about 5.5 million square kilometers or 2.12 million square miles. Sixty percent of this forest lies in Brazil and so this section covers about 3.55 million square kilometers or 1.37 million square miles. Such numbers are difficult to appreciate so perhaps it is more useful to say that the rain forest in Brazil is about the same size as the UK plus France, Germany, Finland, Norway, Sweden, Denmark, Italy, Spain, Portugal, Greece, Belgium, The Netherlands, Austria and Ireland. That is, all of the EC countries plus Norway. It is perhaps interesting to note that most of these countries were heavily forested 1000 years ago. Deforestation there resulted in excellent agricultural land which eventually supported the Industrial Revolution and still remains productive.

To study such a region one requires some data. Fortunately Estaquio Reis, an economist working in Rio de Janeiro, had gathered a panel data set, based on over 300 sub-regions or municipalities for four time periods five years apart. Values for well over a hundred variables had been obtained for each region and year, giving about two hundred thousand individual items of data or numbers. The variables included summaries of local employment, income, agricultural production and prices, population level and change, forestry output and prices, mining production for various materials and metals, and land use estimates for various categories. For example the land use could be divided into original forest, regrown forest, fallow land which may include wet-land, land used for growing crops, and, finally, land used for grazing animals. Once land is used for agriculture it depreciates and one land use naturally converts into another, as will be analyzed later.

Many of the variables being discussed are difficult to define and therefore particularly difficult to measure. Consider whether or not a particular piece of land should be classified as "forested." It may well contain a number of groups of trees with shrubs in between but exactly what quantity and configuration of trees constitutes a forest? Even if each local statistician has a sensible and consistent way of making a decision, remember that in each of the over three hundred municipalities a different person is measuring many of the critical variables and also that very likely different people will be involved when the next set of measurements are taken five years later. The quality of the data is likely to be quite low when different people are making measurements on quantities that are difficult to define and often over very large regions for which travel is difficult and resources available for data collection are limited. The municipalities vary greatly in size; a few are larger than the UK but many are smaller than a typical old-fashioned county. This produces another potential problem in that if there is a serious mis-recording of some number, then a statistician sitting in Brasilia compiling the data from all the regions may not recognize the error because he or she may not have very strong priors about what to expect. This would not be true of interest rates, for example. If most regions were reporting rates around 8 percent but one region reported 80 percent it would clearly be a reporting error. However values for the area that had been deforested in the last five years would be less easy to interpret. Although most of the data we used in our study appeared to be of good quality there was also plenty of evidence that not all of the two hundred thousand pieces of data were perfect. The reaction is that so-called robust methods of analysis have to be used and then interpretation of the results conducted with some care.

The ground-based measures of land use, and thus of deforestation, can be compared to measures obtained from satellite imaging. One of the benefits of the Cold War was the development of very high quality satellite spy photography. Satellites will occasionally pass over the region and accurate photos taken which are then interpreted and a measure of the amount of forest remaining in a region can be obtained. Potentially this is an accurate measure but, being a rain forest, often large regions are obscured by clouds. The ground and space measurements usually agree quite closely but not on all occasions. An example is when a forest edges up to fallow land, the edge consisting of small thin trees perhaps over a wide range. The two methods of measurement are inclined to place the edge of the forest at different places. It was found that the values from space are often helpful for checking the ground-based estimates.

When starting to build a model it is certainly useful to begin with a theoretical foundation but it is important that the theory be relevant. For deforestation an obvious starting point, it would seem, would be consideration of supply and demand. The supply side is fairly easy; it is fixed in the short and medium term as the standing timber and this can be measured, but with some difficulty and thus error. However, the demand side requires a little more thought. The reason for there being a demand for wood varies around the world, apart from the special hard woods used for expensive furniture which are becoming difficult to find everywhere. In parts of Africa and Southeast Asia wood is largely used as a fuel, for heating and cooking, whereas in the more developed countries the major uses are for construction, furniture and paper. Taking either viewpoint it would be quite easy to specify a demand equation using explanatory variables such as timber prices, population size, income levels, and so forth. Although sensible, such an equation would be quite inappropriate for the Amazon region of Brazil where deforestation occurs not because of demand for wood but due to demand for the land on which the trees stand. Poor farmers cut down the forest, burn it, clear the land, and use it to grow crops. This oversimplified account suggests that the correct starting point is the use of supply and demand but applied to cleared land rather than to wood. A demand equation for this land would perhaps concentrate on new farmers in the region, such as immigrants, but again the actual process being studied is rather more complicated. The newly cleared land that is being planted with crops is of poor quality and deteriorates rather quickly until it can no longer be used economically for growing crops but can be used for grazing cattle and sheep. Sometimes the land deteriorates even further and becomes fallow and is possibly of little agricultural value. This discarded land may become forest again over a very long period but over the time span considered by the data this possibility is not relevant. The transition from long-term forest through deforestation to crop land, pasture, and on to long-term fallow land is quite different to

Table 1.1

	$NEWCLEAR_t$	CRP_{t-1}	PAS_{t-1}	FAL_{t-1}
CRP_t	0.10	0.66	0	0.08
PAS_t	0.30	0	1.0	0.11
FAL_t	0.60	0.34	0	0.81

that experienced in Northern Europe or in North America where forests were cut down centuries ago to reveal rich land that has been used successfully ever since, with careful husbandry, and mostly for crops.

For a typical piece of land one has the transition matrix shown in table 1.1. There are three land uses shown: CRP is crop, PAS is pasture, FAL is fallow and NEWCLEAR is land that has been cleared from forest since the previous survey five year ago. The values shown are all essentially transition probabilities and so are quantities between 0 and 1. Thus if for a region $CRP_{t-1} | FAL_t$ is 0.34 it means that 34 percent of the ex-crop land is fallow. NEWCLEAR is the land at time t that did not exist at time $t-1$. Fallow occasionally is included in agricultural land as some of it may have use, either directly or is being held out for later use. The values shown in the table are rather idealized for ease of presentation. It will be seen that they add to one down the columns as each piece of land in some category at time $t-1$ has to be re-allocated at time t. If we take the figures at their face value there seems to be two types of land: that which is suitable for pasture and that which is used for crops.

Old pasture land becomes current pasture land, which is supplemented by 30 percent of the newly cleared land. Thus land classified as pasture will be increasing in quantity, although not necessarily in quality, and much of the land produced by deforestation, within five years, has become pasture, possibly having briefly passed through a crops use stage. Thirty four percent of the old crop land is lost to the fallow category within the five year span, the remaining 66 percent staying as crop land. Current crop land is made up of new cleared land, old crop land, and some old fallow land. Current fallow land includes 60 percent of the land that was newly cleared by deforestation over the last five years, a remarkably large percentage, plus 34 percent of the old crop land but it largely consists of previous fallow land. The amount of fallow land is obviously increasing.

The speed with which newly cleared land becomes pasture, and particularly fallow land over just a five year period and the relatively small contribution it makes to enhancing the stock of crop land suggests that clearing the forest is not an efficient use of the resource. One criticism is that much of the deforestation is done by the cattle industry. The increase in pasture supports this. The figures given in the table are not exact and will vary somewhat by using different estimation methods and different sub-regions of Amazonia but the message remains the same. For the farmers and peasants already in the region to continue crop production at current levels extensive deforestation is needed. Thus deforestation will occur without any new immigrants due just to soil depreciation and current methods of land use. The transition matrix in table 1.1 can only be produced from a panel survey of many districts over several time periods, its formulation and estimation involves little economics but the values obtained suggest relevant economic questions to ask and possible specifications for explanatory models. The initial objective is to quantify the deforestation process and the long-term implications of allowing this process to continue. It may then be possible to consider policies that would slow the destruction of trees in this rain forest, which is considered to be a world resource of considerable value. One such policy would be to somehow boost non-farm employment in the region whilst at the same time importing cheap crops and food. It would certainly be possible to model agricultural production in a typical region and to measure the reaction of farmers to changes in food prices and policies. However, to examine the long-run implications for deforestation it is most appropriate to directly form a demand equation for newly cleared land.

If the variable to be modeled is NEWCLEAR there are many possible explanatory variables that can be considered including various local population measures, previous clearing values, clearing levels in neighboring regions, income per capital, road length, river length, distance to federal and state capitals, area of municipality, and policy variables. The Brazilian government in some periods and locations was encouraging agricultural settlement, largely by designating certain regions as "growth poles" which enjoyed extra favorable investment and tax conditions and by issuing credit incentives granted by the Superintendency for Amazon Development (SUDAM) to some regions. It would be quite easy to write down a sensible model specification, estimate the coefficients by any of several available techniques and then to present the results for interpretation. There are, however, a number of possible pitfalls and problems with such a seemingly simple approach. Some of these are discussed in the following sections.

Model building

For the moment I want to step away from the deforestation study and discuss some of the general questions that arise when building an empirical model. It is useful to propose the existence of a data-generating mechanism to explain the fact that new data are generated almost continuously. As explained earlier, the true mechanism will be extremely complicated and its information will be transformed in many ways from the outcomes of decisions made by economic agents to the data that eventually appear out of the system. One has to hope that these data capture the essential features of the economy and that an empirical model will provide a useful approximation to the generating mechanism of these data.

The classical approach to constructing a model starts with a sound, internally consistent, economic theory which provides a tight specification for the empirical model. This model is then estimated and interpreted. Unfortunately, this strategy towards modeling has not always proved to be a successful one. Models produced in this way often do not fit the data in various important directions. As one pair of applied economists put it, "a recurring problem in empirical studies of consumer and producer behavior is that the regularity properties implied by microeconomic theory have more often than not been rejected" (Rezili and Ozanne (1997)), who then go on to say "such rejection means that empirical work loses a good deal of its theoretical credibility." They point out that a major problem is "the failure of static equilibrium theory to account for dynamic aspects of consumer and producer behavior" and show how the introduction of dynamics into an equilibrium model, by use of a structure known as an error-correcting model, leads to clear improvements. In macroeconomics there has often been a problem relating a theory with the dynamics that might be associated with disequilibrium. The theory often fails to capture vital features of the data, such as trends or seasonal components or some of the structural breaks.

The question of how to use economic theory when constructing an empirical model has generated considerable controversy. One can find advocates at both extremes, some saying that theory contains the only pure truth and so has to be the basis of the model, even to claiming that all "residuals" have to have theoretical explanations, leaving little place for stochastics, uncertainty, or exogenous shocks to the system. At the other extreme, some econometricians have thrown up their hands in despair when trying to find a use for theory and so build "atheoretical" models based just on examination of the data and using any apparent regularities and relationships found in it. Most applied economists take a middle ground, using theory to provide an initial specification and then data exploration techniques to extend or refine the starting model, leading to a form that better represents the data. A formal way to link economic theory and empirical econometrics has been the central aim of a research program by a Norwegian economist, Bernt Stigum, discussed in a lengthy book (1990) and several subsequent papers (e.g. 1995). He discusses how most economic theory is unsuitable for direct empirical use and in what manner the theorist and the econometrician should collaborate to produce a "bridge" from the pristine theory to the more pragmatic data analysis. As a simple example, a theory may relate a variable at time t to another variable at time $t-1$, without any indication of the physical length of the time interval involved: is it a minute, a day, a week, or a month? The answer to this question is very important to the econometrician translating the theory into a practical model.

It is generally accepted that basing a model on a correct theory is good practice, but how do we know if a theory is correct? That is partly my topic for chapter 2. If there are several competing theories they will lead to alternative models which is anathema to anyone taking a scientific viewpoint. However, if we consider economic modelers as producers of commodities and the users of the models as consumers, it is worth remembering that consumers usually prefer having a choice. It is rare for one model to be superior for all possible purposes: forecasting, policy making, conditional forecasts, testing hypotheses, or investigating the effects of a previous policy change, for example. Different users will have different tastes, beliefs, and needs, and will prefer certain types of models. Clearly model providers will not only have to produce the models but also relevant summary statistics allowing model consumers to make sensible choices between them. It is not enough to declare “my model is good” but you should also be expected to prove it!

If no theoretical basis is used and if a complex modeling situation is considered, with many possible explanatory variables and plenty of data, the possibility of “data-mining” or “data-snooping” becomes a problem, particularly now that computing is both fast and cheap. Clearly evaluation procedures need to be applied using data sets that were not involved in the model selection and estimation process, either “out-of-sample” in cross-section or panel analysis, or “post-sample” in time series. It is not sufficient to merely show a statistic that indicates that one model performs better than others; a correct hypothesis test is required, but this takes us into rather technical areas.

One of the basic properties of a model is that it should “balance,” in fact every equation in a system should balance.

To illustrate this idea, consider a single, linear equation with two explanatory variables

$$Y_t = a + b_1 X_{1t} + b_2 X_{2t} + e_t.$$

Suppose from looking at plots of the variables, or by "pre-testing," it is found that some contain a dominating statistical property but that others do not, then these properties have to balance on the two sides of the equation. It is assumed that e_t does not contain the property. Thus, for example, if Y_t contains a clear trend but neither X variable does, the explanatory part of the model cannot explain an important component of the dependent variable Y; and so the equation will be unbalanced. If Y_t and X_{2t} had no trends but X_{1t} had a trend, the equation would balance only if $b_1 = 0$, so now the property determines part of the specification. The same comments would apply if the dominant property was a strong seasonal or a unit root, sometimes called a long-memory or persistence property because the effect of a shock on the economy is noticeable for a long time. The balance question becomes a little more subtle if Y_t does not have the dominant property but both X_{1t}, X_{2t} do have it. It is possible, although somewhat unlikely, that a linear combination of X_{1t}, X_{2t} does not have the property and so the equation can balance. For example, both Xs could contain seasonals, but they might just match so that $b_1 X_{1t} + b_2 X_{2t}$ has no seasonal. The various possibilities are shown in Table 1.2.

If the dominant property is the unit root, or persistence, the case where a pair of Xs have the property but a linear

Table 1.2

P = has dominant property — B = equation balanced

NP = does not have this property — NB = equation not balanced

		X_1, X_2 both P	X_1, X_2 both NP	X_1, X_2 one P, one NP
Y	P	B	NB	B
	NP	NB?	B	NB

Note: NB? Will be NB unless there exists a linear combination of Xs which is NP.

$$Y_t = b_1 X_{1t} + b_2 X_{2t} + \text{error} - \text{(assumes error is } NP\text{)}$$

combination does not have it is found to occur quite frequently in economic time series, and is given the name "cointegration."

There can be a problem with a balanced set of variables on some occasions when an apparent relationship is found when truly none is present. This is well known to occur between a pair of smooth, or persistent, time series that are actually independent of each other but are modeled using standard techniques, in which case so-called "spurious relationships" can occur. Once one is aware of the possibility, appropriate techniques can be employed to avoid the difficulty.

A similar difficulty can occur with panel data. Suppose that X_{jt}, Y_{jt} are series measured at time t for the jth region, that each depend heavily on some measure of the "size" of